Detecting Ironic Speech Acts in Multilevel Annotated German Web Comments

Bianka Trevisan[1], **Melanie Neunerdt**[2], **Tim Hemig**[1], **Eva-Maria Jakobs**[1], **Rudolf Mathar**[2]
Textlinguistics and Technical Communication[1],
Institute for Theoretical Information Technology[2],
RWTH Aachen University, Germany

Abstract

Ironic speech act detection is indispensable for automatic opinion mining. This paper presents a pattern-based approach for the detection of ironic speech acts in German Web comments. The approach is based on a multilevel annotation model. Based on a gold standard corpus with labeled ironic sentences, multilevel patterns are determined according to statistical and linguistic analysis. The extracted patterns serve to detect ironic speech acts in a Web comment test corpus. Automatic detection and inter-annotator results achieved by human annotators show that the detection of ironic sentences is a challenging task. However, we show that it is possible to automatically detect ironic sentences with relatively high precision up to 63%. [1]

1 Introduction

Automatic detection of irony in text is a challenging task. However, typical characteristics, e.g., emoticons, inherent in Web comments, are strong indicators for ironic speech acts. This forms a new basis for the detection of irony. In this paper, we present a pattern-based approach for the detection of ironic speech acts in German Web comments. Challenges in the identification of ironic speech acts concern the fact that the identification of irony without the context is almost impossible (Sandig, 2006). Hence, sophisticated techniques are required that allow for irony detection (Mihalcea and Strapparava, 2006). For Web comments, however, typical characteristics or indicators of ironic speech acts are identified such as winking emoticons (Neunerdt et al., 2012), quotation marks, positive interjections (Carvalho et al., 2009) or opinionated words (Klenner, 2009). In contrast to standardized texts, we believe that in Web comments such characteristics allow for better detection of ironic speech acts. Nevertheless, the question is, can ironic speech acts reliably and automatically be detected based on these indicators in Web comments and what challenges arise?

Contrary to the common conceptualization, we assume that ironic speech acts are not only characterized by features at the text surface but rather by a whole set of linguistic means whose specific combination (*pattern*) indicates a specific speech act such as *IRONIZE*. In order to identify and define these patterns, we suggest a fine-grained multilevel annotation model where different linguistic means are considered. The annotation on different levels allows for level-vise and level-combined pattern analysis. The proposed approach works as follows.

First, based on a gold standard Web comment corpus typical ironic multilevel patterns (training patterns) are determined according to statistical and linguistic analysis for the detection of ironic speech acts. The gold standard corpus is manually annotated on all annotation levels. Second, the revealed training patterns serve to detect ironic speech acts in a huge Web comment test corpus. The test corpus is tokenized and Part-of-Speech

(POS) tagged automatically by the *WebTagger* proposed in (Neunerdt et al., 2013a). Based on the tokens and POS tags, the Web comments are labeled on multiple annotation levels by the *AutoAnnotator* (Trevisan et al., 2014). Detection results achieved with the training patterns are manually annotated by different annotators and evaluated.

The paper is structured as follows. Section 2 summarizes related work on irony conceptualization and detection. In Section 3, we introduce the multilevel annotation scheme and the pattern detection method. Section 4 reports the different corpora and experimental results. They are discussed in Section 5. In Section 6 we conclude our work and outline future work.

2 Related Work

In linguistics, there is a huge research regarding speech act theory. In our work, we follow the approach of (Sandig, 1979) who focuses on specific speech acts, namely evaluative speech acts such as ironic speech acts (*linguistic evaluation theory*). (Sandig, 1979), and in the following (Ripfel, 1987), conceptualizes the process of evaluation, respectively, an evaluative speech act as an act in which a subject evaluates an object with a specific purpose using evaluative expressions or linguistic means such as idiomatic expressions (e.g. *Too many cooks spoil the broth*), attributes (e.g. *right* vs. *wrong*) or evaluative lexis (e.g. *brick*) (Trevisan and Jakobs, 2010; Trevisan, 2014). The linguistic means can be used for different evaluative purposes, such as stylistic and pragmatic means for the purpose of *addressee-oriented evaluation*. In this kind of evaluation, the speaker formulates and modifies speech acts according to the evaluative intention of the communication situation and the addressee. The modification of the speech act is done by changing the style or manner of formulation. Possible speech acts are, for example, *IRONIZE*, *STRENGTHEN*, or *WEAKEN*.

Thereby, irony is an extremely complex or form-rich speech act, exemplified by the fact that multiple linguistic means are used for different phenomena, such as argument something ad absurdum, reverse something, or explicate logical relationships too clearly (Bohnes, 1997). In addition, challenges in the detection of ironic speech acts relate, particularly, to the strong interpretive ductus and context-dependency. Hence, regarding the focus of this paper, the automated detection of ironic speech acts in Web comments, the challenging task is to deal with different forms of irony and to find out which indicators are most useful for irony detection.

In computational linguistics, there is initial work done regarding the automated detection of irony in text. Approaches in this context mainly focus on the identification of emotions or humor. (Carvalho et al., 2009) identified surface clues of positive ironic sentences in comments applying a rule-based approach. In this approach, patterns are defined whose occurrence shows evidence of certain surface clues, e.g., the pattern *(ADJpos|Npos)* as indicator for irony by quotation marks. The authors found out that irony-indicating surface characteristics in sentences with a positive predicate are besides quotation marks, onomatopoetic expressions, heavy punctuation marks, and positive interjections. (Mihalcea and Strapparava, 2006) used automatic classification techniques to identify humour in one-liners, i.e., short sentence characterized by simple syntax, use of rhetoric means (e.g. alliteration), and creative language constructions. The results show that it is possible to distinguish humorous and non-humorous sentences, but the technique failed regarding the automatic and reliable identification of irony. Therefore, more sophisticated techniques are needed.

Beyond the reported approaches, there are several more in computational linguistics that provide hints on indicators of ironic speech acts in different text types. For instance, winking emoticons (*;)* and *;-)*) are irony indicators especially in chat communication (Beißwenger et al., 2012) and Web comments (Neunerdt et al., 2012). (Klenner, 2009) points out that in prose texts a positive attributive adjective and a negative noun ($ADJA^{+}$ NN^{-}) indicate an ironic speech act.

However, all described approaches do not provide a full-automated solution for the detection of ironic speech acts.

3 Methodological Approach

To detect ironic speech acts in Web comments, different indicators of multiple linguistic levels

are considered and subsumend into patterns. The multilevel annotation is described in Section 3.1, the methodology for pattern-based detection of ironic speech acts in Section 3.2.

3.1 Multilevel Annotation

In order to define patterns for detection, a linguistic multilevel annotation model proposed by (Trevisan, 2014) is applied. In the model, Netspeak-specific pecularities are considered and modeled such as non-standard parts of speech (e.g. Leetspeak), interaction signs (e.g. emoticons), different speech acts (e.g. *IRONIZE*) or syntactic peculiarities of Web language such as missing punctuation marks (Trevisan, 2014). Totally, the model contains seven linguistic annotation levels (graphematic, morphological, syntactic, semantic, pragmatic and polarity level, level of rhetorical means) and its sub-levels. At each level, different linguistic means are annotated, for instance, at the *pragmatic or target level* 30 different speech acts. The annotation model is based on the assumption that the annotated linguistic means and levels provide evidence or clues for the detection of evaluative speech acts in Web comments.

In this approach, we particularly consider *ironic speech acts* as target class. For the detection of ironic speech acts, three annotation levels out of seven are selected: POS level, graphematic level, and token polarity level. These levels are chosen due to the fact that a tool exists to annotate such levels automatically (*AutoAnnotator*) (Trevisan et al., 2014). We assume that indicators of these automatically annotated levels are mutually dependent in their appearance and, thus, in combination turn into patterns that can be more or less reliably used for the automatic detection of ironic speech acts. As speech act boundaries, we consider the beginning and the end of a sentence, determined by the corresponding POS tag on POS level.

Hereafter, the annotation levels used for pattern creation are described briefly in chronological order. Note that the terms label and tag are used synonymously.

- *Level 1 - POS level* (l_1): At the POS level, to each token a morphosyntactic category is assigned providing information about part of speech and syntactic function. POS tags are assigned according to the Stuttgart-Tuebingen Tagset (STTS), and lemma information according to a special lexicon (Schmid, 1995); (Schiller et al., 1999). In total, the tagset consists of 54 tags. Since the tagset was developed on standard texts such as newspaper articles, tag correspondences had to be defined for Netspeak-specific expressions such as emoticons (EMO = $.) (Trevisan et al., 2012); (Neunerdt et al., 2013b).

- *Level 2- Graphematic level* (l_2): At the graphematic level, expressions at the text surface as well as grapho-stilistic features that show special notational styles are annotated following (Gimpel et al., 2011). In total, eight labels are distinguished: adressing terms (e.g. *@[John], 2[heise]*; label: ADD), words with capital letters within (e.g. *CrazyChicks*; label: BMAJ), emoticons (e.g. *;-)* ; label: EMO), iterations (e.g.*yeeeeeees* ; label: ITER), leetspeak (e.g. *W1k1pedia* ; label: LEET), words in capital letters (e.g. *GREAT*; label: MAJ), markings (e.g. **[quiet]**; label: MARK) and mathematical symbols (e.g. +; label: MAT).

- *Level 3 - Token polarity level* (l_3): At the level of token polarity, the polarities of individual tokens are annotated, i.e., the polarity of words or interactive signs. There are five categories distinguished: negative token (e.g. *harmful*; label: $-$), positive token (e.g. *suitable*; label: $+$), deminisher (e.g. *less*; label: $\%$), intensifier (e.g. *much*; label: $\wedge$) and reverser (e.g. *not*; label: $\sim$).

3.2 Pattern-based Detection

The goal of our work is to detect ironic speech acts in Web comments. The overall approach is simple, based on statistical and linguistic criteria. Training patterns are defined based on a gold standard corpus, which are later used to detect sentences representing ironic speech acts (*ironic sentences*) in a Web comment corpus. In the following, we mathematically describe the two steps of our approach: First, we describe the identification

of frequent patterns over multiple annotation levels in the gold standard corpus and, second, the search process of the defined patterns for the detection of ironic speech acts in the test corpus. Therefore, we consider the gold standard corpus consisting of K sentences with labeled ironic sentences. Note that the sentence boundaries are determined by the corresponding POS tag information. Each sentence $k \in K$ contains a sequence of N_k tokens:

$$(w_1, \ldots, w_{N_k}) \in \mathcal{W}^{N_k}$$

where $\mathcal{W}$ contains all possible tokens. For each annotation level $l = 1, \ldots, L$, the corresponding labels

$$\left(t_1^l, \ldots, t_{N_k}^l\right) \in \left(\mathcal{T}_l \cup \{\epsilon\}\right)^{N_k}$$

are assigned, where $\mathcal{T}_l$ represents the set of L_l labels for a particular annotation level l:

$$\mathcal{T}_l = \left\{c_1^l, \ldots, c_{L_l}^l\right\}.$$

In our approach, we consider $L = 3$ levels, e.g., the token polarity level with $\mathcal{T}_3 = \{+, -, \%, \wedge, \sim\}$ as described in Section 3.1. Note that on some levels it is not mandatory to annotate each token. Hence, tokens which are not annotated are labeled with ϵ. The gold standard corpus labels are assigned manually by human annotators. The test corpus is labeled by means of *AutoAnnotator*, which is described in Section 3.1.

In order to determine frequent patterns in the gold standard, we first determine the label combinations of a sentence. First, for each level a feature vector

$$\mathbf{m}^l = \left(m_1^l, \ldots m_{L_l}^l\right) \tag{1}$$

with

$$m_p^l = \begin{cases} 1 & \exists\, n : t_n^l = c_p^l \\ 0 & elsewise \end{cases}$$

is calculated. The single components m_p^l indicate the presence (1) or absence (0) of a particular label c_p^l. These feature vectors are determined for all sentences $k \in \mathcal{K}$ as $\mathbf{m}_k^l$. Exemplarily, for the sentence k: *"Schon mal zu optimistisch an ein Projekt ran gegangen ;o)?" ("Have you ever tackled a project too optimistic ;o?")*, the feature vector, e.g., for level 3, results in $\mathbf{m}_k^3 = (1, 0, 0, 1, 0)$.

In order to detect statistical peculiarities, we determine the frequency of all occurring label combinations for single level, tuples and triples of levels, i.e., for n levels $l_1, \ldots, l_n \in \{1, \ldots, L\}$ and jointly occurring feature vectors $\mathbf{m}^{l_1}, \ldots, \mathbf{m}^{l_n}$ we calculate

$$N(M^{\mathbf{P}}) = \left|\left\{k \in \mathcal{K} \mid \mathbf{m}_k^{l_i} = \mathbf{m}^{l_i}, \forall i = 1, \ldots, n\right\}\right|$$

with

$$\mathbf{P} = \{l_1, \ldots, l_n\}$$

and

$$M^{\mathbf{P}} = \left(\mathbf{m}^{l_1}, \ldots, \mathbf{m}^{l_n}\right).$$

Tuples and triples are in the following sorted according to their frequencies. Example tuples and triples are given in the forth column of Table 1. According to the top frequencies and considering the pattern frequency in ironic speech acts (*IRONIZE*) only $N_I(M^{\mathbf{P}})$ compared to their frequency in other speech acts a set of tuples and triples is selected. The selected patterns fullfill $N_I(M^{\mathbf{P}})/N(M^{\mathbf{P}}) \geq 0.8$ and serve for further linguistic analysis. Based on the qualitative results, some tuples and triples are slightly modified or added due to the results, see Section 4.

The extracted tuples and triples serve to detect ironic sentences in a test corpus. The test on an arbitrary sentence works as follows. First, we calculate its feature vectors M_t according to (1). A sentence t is declared ironic if one of the defined training patterns $M^{\mathbf{P}}$ fulfills the equation

$$\text{IRONIC}(M^{\mathbf{P}}, M_t) = \prod_{l \in \mathbf{P}} I(\mathbf{m}^l, \mathbf{m}_t^l) = 1$$

with

$$I(\mathbf{m}^l, \mathbf{m}_t^l) = \prod_{p=1,\ldots,L_l} IM(m_p^l, m_{t,p}^l),$$

i.e., on each level $l \in \mathbf{P}$ at least the labels seen in the training pattern have to be present. Hence, we define

$$\text{IM}(m_p^l, m_{t,p}^l) = \begin{cases} 1 & m_p^l \leq m_{t,p}^l \\ 0 & elsewise. \end{cases}$$

We use the minimum criteria fit instead of an exact match in order to relax the restrictions. For example, on the POS annotation level an exact pattern match would lead to very strong restrictions.

4 Experimental Results

The aim of our paper is the identification of indicators and patterns that allow reliable automatic detection of ironic speech acts in Web comments. To this end, we first search for indicators of ironic speech acts in a multilevel annotated gold standard corpus (Section 4.1). In a second step, the extracted patterns are used to detect ironic speech acts in the Web comment test corpus and extract the corresponding sentences (Section 4.2).

4.1 Corpora

As an exemplary corpus, a topic-specific Web comment corpus is collected from *Heise.de*, which is a popular German newsticker site treating different technological topics. Web comments from 2008 and 2009 are collected. In total, the *Heise* corpus contains approximately 15 Million tokens.

For training purposes, a small corpus *HeiseTrain* containing Web comments with approximately 36,000 tokens is separated according to different criteria. The remaining Web comments serve as test corpus (*HeiseTest*) to evaluate the sentence extraction according to patterns for ironic speech acts (see Section 3.2). *HeiseTrain* serves as gold standard, which is manually annotated on multiple levels according to Section 3.1, among others the target level with labeled ironic sentences. For manual multilevel annotation, the tool *EXMARaLDA* is used, which is formally applied for conversational research, e.g., the analysis of audio transcripts. The annotation is performed by five annotators (Trevisan, 2014). Annotator 1-4 annotate at all levels the entire corpus. Annotator 5 annotates only those text segments, where no majority decision could be determined between annotator 1-4. Finally, the gold standard is derived from the annotation of annotator 1-5.

Figure 1 shows the corpus statistics for the target level on which evaluative speech acts are annotated. Additionally, l_1 (POS level), l_2 (graphematic level) and l_3 (token polarity level) statistics are given for the 220 ironic speech acts (*IRONIZE*) exclusively. As evident from the statistics for target level, the top 5 ranked speech acts reach more than half of all identified speech acts. Therein, the speech act *IRONIZE* (n=220)

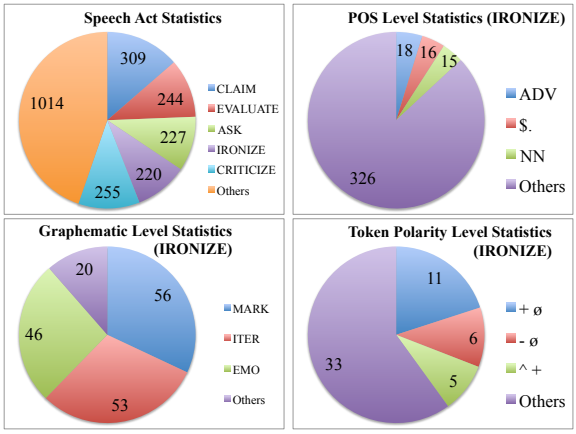

Figure 1: *HeiseTrain* corpus statistics on the target level and different annotation levels.

is ranked in the top 5 of the most often occurring speech acts in *HeiseTrain*. Second, on l_1 the most occuring tags are ADV (n=18), $. (n=16) and NN (n=15). An outstanding result is obtained for l_2: almost 90% of the most identified graphematic labels are the indicators MARK (n=56), ITER (n=53) and EMO (n=46). As most relevant patterns for token polarity, the combination of a positive token (+) and a non-valuing token (ø) are identified (n=11).

For the *HeiseTest* corpus, the multilevel annotation is carried out automatically. The POS tagging is performed by means of *WebTagger* (Neunerdt et al., 2013b) whereas level 2 and 3 as well as the basic level are annotated by means of the multilevel annotation tool *AutoAnnotator* (Trevisan et al., 2014). The *AutoAnnotator* is a rule-based and lexicon-based annotation system and uses the EXMARaLDA editor as data format. Besides POS tagging accuracies of about 95%, accuracies on other levels have to be examined in more detail.

4.2 Ironic Speech Act Patterns

Initially, multilevel patterns are determined according to the method described in 3.2 based on the *HeiseTrain* corpus. As a result of statistical evaluations, we analyze three statistical patterns with patterns over the levels l_1, l_2 and l_3. Results are depicted in the first three rows of Table1 marked as type *STAT*. The statistical pattern serve as basis for the derivation of further patterns that are modeled based on linguistic assumptions

and involve features that have been identified in previous studies, see Section 2. To be precise, we integrate the indicators l_3:(+, -) claimed by (Klenner, 2009) as well as the indicators quotation marks l_2:(*MARK*) and laughter expression l_2:(*EMO*) of (Carvalho et al., 2009). In conclusion, we obtain a type of pattern which is composed primarily of the statistical pattern and completed by additional features (type: *STAT+LING*, e.g., $P_{SL}1_{ITER}$ = P_S1_{ITER} added by l_3: "-") as well as a type of pattern that contains only linguistically motivated, non-statistical features (type: *LING*). Finally, nine patterns with features originate from two or three different levels (tuple:$|\mathbf{P}| = 2$, triple:$|\mathbf{P}| = 3$) are used and analyzed for the detection of ironic speech acts. All patterns and some *HeiseTrain* and *HeiseTest* corpus statistics are depicted in Table 1. Column five $N(M^{\mathbf{P}})$ depicts the number of exact pattern matches in the *HeiseTrain* corpus. Furthermore, the number of detected sentences with our method based on a minimum criteria fit described in 3 is given in column 6 for the gold standard corpus *HeiseTrain* (#Matches *GS*) and in column 7 for the *HeiseTest* corpus (#Matches *HT*). Finally, the occurrence of each pattern in the *HeiseTest* corpus (#Matches *HT*) is determined. The sentences with pattern matches in the *HeiseTest* corpus are extracted for pattern evaluation (see Table 2).

As evident from Table 1, the statistically determined pattern P_S2_{ITER} achieves most matches in both corpora. Rather few matches provide the linguistic patterns P_L2_{MARK} and P_L3_{MARK}.

In order to assess the usefulness of the patterns for irony detection, the extracted sentences are annotated manually and further evaluated by an inter-annotator agreement study, see Table 2. For each pattern, a set of 200 randomly chosen sentences is evaluated; less sentences are evaluated for the pattern P_L2_{MARK} and P_L3_{MARK}. Two annotators had to decide whether a sentence is an ironic or non-ironic sentence (A1 Ironic vs. A2 Ironic). Thereby, the sentence annotation is performed without considering any context, which is contrary to current methods of irony classification. For instance, (Carvalho et al., 2009) use two more classes for the annotation of unclear cases, e.g., where the context is needed or the decision. In our case, we redesigned this approach for two reasons: First, since the corpus is topic-related and the annotators are very familiar with the data, the consideration of the context can be neglected, mainly. Furthermore, giving a default class for cases, which are not clear, prevents the annotator from a clear decision, i.e., in case of doubt, the annotator would opt for the default class.

Consequently, the inter-annotator agreement between A1 and A2 is calculated (IAA(A1, A2)). In those cases, in which there is no match between A1 and A2, A3 decides whether the sentence is ironic or non-ironic (#Sentences A3). Based on the classification of the annotators, the proportion of sentences is determined that is classified by the majority as ironic. The similarities between the annotators (A1=A3; A2=A3) are listed in the last two columns (see Table 2).

The results of the inter-annotator agreement demonstrate two findings, particularly: Those patterns that brought forth the lowest number of pattern matches in Table 1 reached the best inter-annotator agreement (P_L2_{MARK} = 62.79% and P_L3_{MARK} = 63.63%, see Table 2). At the same time, the pattern that brought forth the highest number of pattern matches in Table 1 reached the lowest inter-annotator agreement (P_S2_{ITER} = 25.34%, see Table 2).

Furthermore, the inter-annotator agreement shows that the correspondence between A1 and A2 and between A2 and A3 has the largest irregularities regarding the linguistic patterns (type: LING). Here, the annotators frequently disagreed whether the examined sentence is an ironic or non-ironic sentence. In contrast, the results for the pattern of type STAT and STAT+LING are much more consistent.

5 Discussion

The results show that particularly those linguistically motivated patterns achieve a high inter-annotator agreement. The pattern with the highest inter-annotator agreement consists of self-selected, linguistic features that are based on assumptions, previous statistical results (see Section 4.1), and that are taken from the literature. However, statistical results serve as starting point for the linguistic motivation of such multilevel patterns. These results suggest two conclusions: First, the gold standard corpus used for statisti-

Pattern	Type	|P|	Patterns $M^{\mathbf{P}}$ (Tuples,Triples)	$N_I(M^{\mathbf{P}})$	#Matches *GS*	#Matches *HT*
P_S1_{ITER}	STAT	3	l_1: (\$., ADJD) l_2: (ITER) l_3: (+)	2	2	2640
P_S2_{ITER}	STAT	2	l_1: (\$., ADV, NN) l_2: (ITER)	4	17	28751
P_S3_{ITJ}	STAT	2	l_1: (\$., ITJ) l_3: (+)	2	6	3368
$P_{SL}1_{ITER}$	STAT+LING	3	l_1: (\$., ADJD) l_2: (ITER) l_3: (+, -)	0	1	421
$P_{SL}2_{ITER}$	STAT+LING	3	l_1: (\$., ADV, NN) l_2: (ITER) l_3: (+, -)	0	0	422
$P_{SL}3_{ITJ}$	STAT+LING	2	l_1: (\$., ITJ) l_3: (+, -)	1	1	549
P_L1_{MARK}	LING	3	l_1: (NN) l_2: (MARK) l_3: (+, -)	0	0	826
P_L2_{MARK}	LING	3	l_1: (ITJ) l_2: (MARK) l_3: (+, -)	0	0	43
P_L3_{MARK}	LING	2	l_2: (EMO, MARK) l_3: (+, -)	1	1	22

Table 1: Extracted patterns and their corpus frequencies in *HeiseTrain*. Explanation: P=pattern, S=statistical pattern, L=linguistic pattern, SL=statistical, linguistic pattern, ITER=iteration, MARK=marking, ITJ=interjection, P=number of pattern-inherent levels, $M^{\mathbf{P}}$=pattern, $N_I(M^{\mathbf{P}})$=exact pattern frequency in *IRONIZE* of *HeiseTrain*, #Matches *GS*=minimum citeria fit pattern frequency in *IRONIZE* of *HeiseTrain*, #Matches *HT*=minimum citeria fit pattern frequency in *HeiseTest*.

Pattern	A1 Ironic	A2 Ironic	IAA(A1,A2)	#Sent. A3	Ironic(A1,A2,A3)	A1=A3	A2=A3
P_S1_{ITER}	29.86%	35.07%	73.93%	55	30.81%	71.09%	63.98%
P_S2_{ITER}	21.72%	34.84%	66.97%	73	25.34%	73.75%	69.68%
P_S3_{ITJ}	27.96%	49.28%	64.45%	75	37.91%	64.45%	58.29%
$P_{SL}1_{ITER}$	25.82%	38.50%	71.36%	61	31.92%	68.54%	67.13%
$P_{SL}2_{ITER}$	27.11%	51.11%	65.33%	78	37.33%	62.67%	59.11%
$P_{SL}3_{ITJ}$	25.46%	47.22%	69.00%	67	33.80%	62.50%	64.81%
P_L1_{MARK}	50.95%	45.71%	70.48%	62	36.49%	53.35%	22.28%
P_L2_{MARK}	44.18%	69.77%	60.47%	17	62.79%	34.88%	51.16%
P_L3_{MARK}	59.09%	45.45%	68.18%	7	63.63%	50.00%	45.45%

Table 2: Results achieved for sample matches in *HeiseTest*. Explanation: A1=annotator 1, A2=annotator 2, A3=annotator 3, IAA=inter-annotator agreement, #Sent.A3=number of sentences annotated by A3, Ironic(A1,A2,A3)=majority decision over all annotators.

cal analysis and pattern definition with a scope of about 36,000 tokens is too small. For future studies, a larger gold standard corpus is recommended. Second, to avoid methodological effects due to the sample, the gold standard corpus, for example, should be compiled due to different selection criteria, e.g., topic or domain.

In addition, comparing the inter-annotator results with those from a previous study, it is evident that the choice of the annotators does alter the result. The annotators who conducted the inter-annotator agreement in this study are all familiar with the subject and the corpus. All three (A1, A2, A3) were involved in the development of the complete annotation scheme. However, previous studies have shown that in particular, a much higher inter-annotator agreement is reached with those annotators who had no prior knowledge regarding the annotation model or topic (Trevisan, 2014). Thus, it should be considered whether future inter-annotator agreement studies are carried out only with new, previously non-involved annotators.

With regard to the investigated pattern, other features should be taken into consideration. In the present study, only the indicators marking (label: MARK), interjection (label: ITJ) and iteration (label: ITER) are considered. A rather small proportion is ascribed to the feature emoticon (label: EMO) in contrast to the literature. Moreover, not considered features concern the semantic level and the morphological level, for example, usage regularities of topic-specific words or word types (e.g. redemptions such as *nen — einen = one*) in ironic sentences.

6 Conclusion and Outlook

In this paper, we presented a method for the automatic identification of ironic speech acts in German Web comments. As a result, ironic sentences were identified by the annotators with an accuracy of up to 63%.

Future work will focus on the iterative extraction and development of primarily linguistic patterns. To be precise, the results of the inter-annotator agreement will be validated in future

studies. Thereby, the immediate context of each sentence will be involved, i.e., the previous and the following sentence will be shown to the annotators. We assume that a higher accuracy will be achieved in the identification of irony. In addition, the investigated corpus will be enlarged in order to obtain a higher sample, identify more patterns also statistically and ensure the methods reliability.

Acknowledgments

We owe gratitude to the Excellence Initiative of the German Federal and State Government as well as Eva Reimer, Julia Ninnemann, and Simon Ruppel for their support in data processing.

References

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2012. A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative*, pages 1 – 31.

Ulla Bohnes. 1997. Compas-b. beschreibung eines forschungsprojektes. magisterarbeit im fach neuere deutsche sprachwissenschaft. Master's thesis, Universität des Saarlandes.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "So Easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 53–56, New York, NY, USA. ACM.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 42–47.

Manfred Klenner. 2009. Süsse Beklommenheit und schmerzvolle Ekstase. Automatische Sentimentanalyse in den Werken von Eduard von Keyserling. *Tagungsband der GSCL-Tagung, Gesellschaft für Sprachtechnologie und Computerlinguistik*, 30(2).

Rada Mihalcea and Carlo Strapparava. 2006. Learning to Laugh (automatically): Computational Models for Humor Recognition. *Computational Intelligence*, 22(2):126–142.

Melanie Neunerdt, Bianka Trevisan, Rudolf Mathar, and Eva-Maria Jakobs. 2012. Detecting Irregularities in Blog Comment Language Affecting POS Tagging Accuracy. *International Journal of Computational Linguistics and Applications*, 3(1):71–88, June.

Melanie Neunerdt, Michael Reyer, and Rudolf Mathar. 2013a. A POS Tagger for Social Media Texts trained on Web Comments. *Polibits*, 48:59–66.

Melanie Neunerdt, Bianka Trevisan, Michael Reyer, and Rudolf Mathar. 2013b. Part-of-Speech Tagging for Social Media Texts. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 139–150, Darmstadt, Germany, September.

Martha Ripfel. 1987. Was heißt bewerten? *Deutsche Sprache*, 15:151–177.

Barbara Sandig. 1979. Ausdrucksmöglichkeiten des bewertens. ein beschreibungsrahmen im zusammenhang eines fiktionalen textes. *Deutsche Sprache*, 7:137–159.

Barbara Sandig. 2006. *Textstilistik des Deutschen*. de Gruyter, Berlin/New York.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. University of Stuttgart.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer.

Bianka Trevisan and Eva-Maria Jakobs. 2010. Talking about mobile communication systems: verbal comments in the web as a source for acceptance research in large-scale technologies. In *Professional Communication Conference (IPCC), 2010 IEEE International*, pages 93–100.

Bianka Trevisan, Melanie Neunerdt, and Eva-Maria Jakobs. 2012. A Multi-level Annotation Model for Fine-grained Opinion Detection in German Blog Comments. In *11th Conference on Natural Language Processing (KONVENS)*, pages 179–188, Vienna, Austria, September.

Bianka Trevisan, Tim Hemig, and Eva-Maria Jakobs. 2014. *AutoAnnotator: A Tool for Automated Multilevel Annotation of Web Comments*. In preparation.

Bianka Trevisan. 2014. *Bewerten in Blogkommentaren. Mehrebenenannoation sprachlichen Bewertens*. RWTH Aachen University.